Improving Visual Question Answering Using Semantic Analysis and Active Learning

CMPT 726: Machine Learning Course Project

SFU SCHOOL OF COMPUTING SCIENCE

1 Introduction

- Visual Question Answering (VQA) is the task of answering questions about a given piece of visual content such as an image, video, or infographic.
- One of the problems that may arise in this task is about the lack of unlabeled dataset resulting in not-generalized models.
- Active learning is an iterative supervised learning solution to the situations in which unlabeled data is abundant, but manually labeling is expensive. In such a scenario, learning algorithms can actively query the user/teacher for labels.
- We aim to overcome the lack of labeled data by introducing a model that will help us understand the features of objects in a picture. We use this model as an oracle for a VQA model, and we define a loss function based on the semantic similarity of the question and the generated caption by the oracle. With these architecture, we can follow a procedure based on active learning methods.

2 Approach and Ideas

In order to overcome the insufficiency of labeled data, we use a captioning model to generate a description of the input the image.
For each image, there are some questions that are asking about a specific detail in image.
To generate a label, we created a neural network model that is able to generate a detailed description of the image. In order to find a correct answer, a model is used to find semantic similarity between the question and the windowed parts of description.

1. By calculating the distance for different placement of the window, a window that has minimum distance with the question is chosen.
2. After finding this window, we now aim to optimize the result of the VQA model by minimizing the distance between the output of the model and the word that has maximum similarity on the windowed part of the description.
3. Intuitively, this will result in the network to guess words that are close in the meaning to the objective of the question.

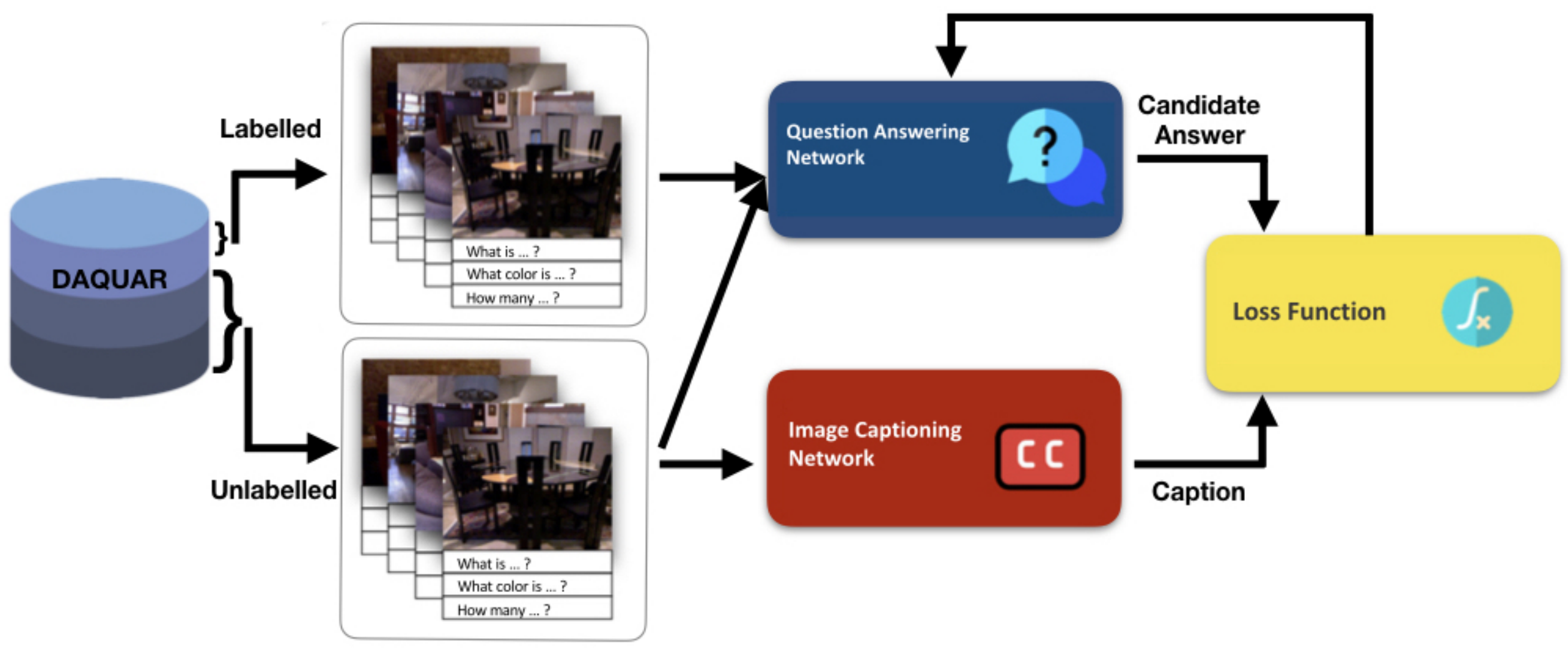

Figure 1: High-level diagram of the problem

3 Model Architecture

Our model consists of three major parts, named VQA model, image caption generator and semantic similarity calculator.

3.1 VQA Model

This part of the model has the responsibility to get a picture and a question as inputs and generate a single word answer. We used a deep neural network model which has three main parts:

1. Pre-trained ResNet 18 model that receives the image part as its input.
2. GRU-based LSTM recurrent neural network that gets the question as its input.
3. These two parts are concatenated together and make the input to the third part, which is a fully connected neural network and produces an output in $\mathbb{R}^{300}$.

We used DAQUAR dataset to train and analyze the performance of this model.

3.2 Image Caption Generator

For this part of the model, we trained a deep neural network independently on the MS-COCO dataset. Then, we used this trained model as an oracle for the VQA model to get the description of each image that we want to give to the VQA model as an input. A description of our model's components is:

1. Extracting the features from the lower convolutional layer of InceptionV3, giving us a three dimensional vector of shape (8, 8, 2048).
2. Squashing previous vector to a shape of (64, 2048).
3. This vector is then passed through a convolutional neural network encoder (which consists of a single fully connected layer).
4. Using a recurrent neural network based on GRU that attends over the image to predict the next word [2].

3.3 Semantic Similarity Calculator

For calculating semantic similarity, we use GloVe [1], which is an unsupervised learning algorithm for obtaining vector representations for words. This algorithm has this essential property that encodes similar words to some vectors that have small Euclidean distance. This feature helps us to both find the optimum window in the question and the most similar word in the optimum window to the output of the VQA model.

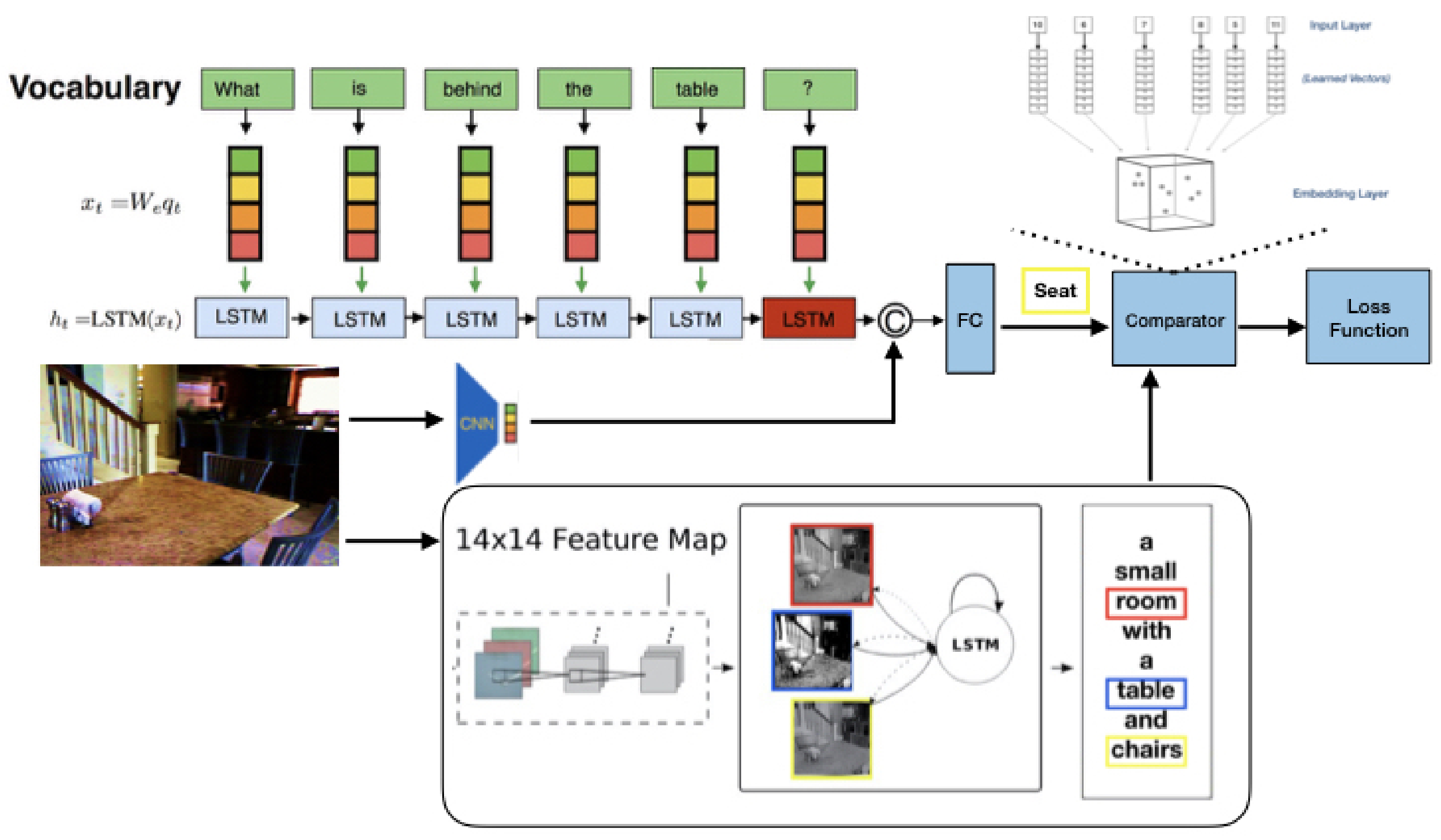

Figure 2: Structure of networks used in the model

Loss Function

For the loss function, we used a cross-entropy measure on the maximum similarity that we can find between the words of the optimum window and the output vector presented by the VQA model. Formulation of this function comes as follows.

$$class = \arg\max_i \|output - optimum\_window[i]\|_2 \tag{1}$$

$$Loss(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) \tag{2}$$

4 Results

Experiments results shown in figure 3 indicate that there is a significant gap between testing and training loss in our approach. The training loss is the result of comparing the predicted output with the candidate labels which the model got from the caption. Since we are training the model on the unlabeled data, the training loss is not a precise measure to evaluate the model. But to compute testing loss, we used labeled test set provided by DAQUAR dataset. This will allow us to compare our results with the classic approach that uses entire training set to train the model. Second chart compares the two testing losses in order to demonstrate the differences between our approach and classic approach. We note that the difference between these two errors are negligible, however, our approach has a better generalization on the unseen data.

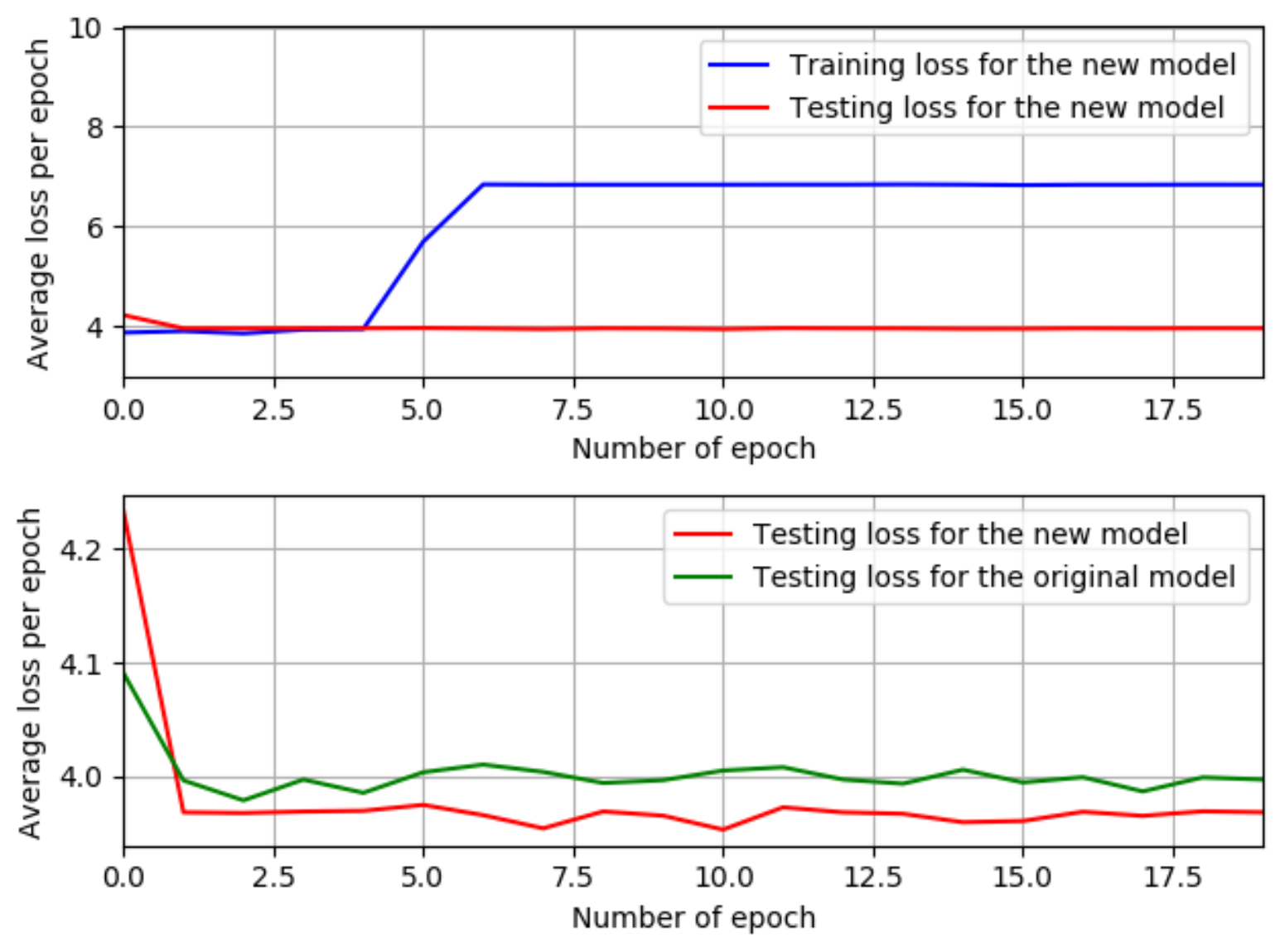

Figure 3: Training and testing loss versus each epoch in the experiment

Also, the average values of testing loss for both approaches are provided in the table 1.

Model	Testing loss
Classic model	4.00232881
New model	3.97967066

Table 1: Comparison between classic approach and new approach

5 Conclusion

In this project, we aimed to train a model for the VQA problem without using a great amount of labeled data. Instead, we used an active learning based approach, and defined a trained image captioning network as an oracle for the labelling task on the pair of questions and images without having the exact answer. To increase the precision, we used a model based on GloVe algorithm that helps us to find most related portions of the caption to the question. The results of this project shows that this approach can result to a model that its loss is in the same standard of classical approaches.

References

[1] J. Pennington et al. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[2] K. Xu et al. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015.